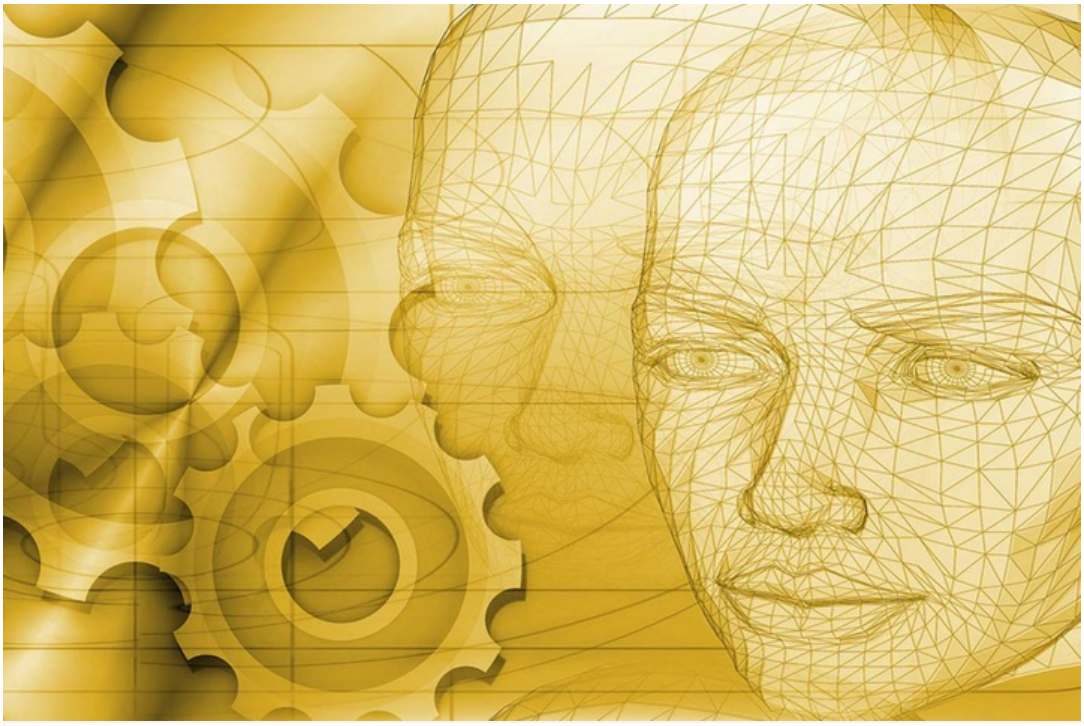
October 9, 2018 | Mike Loukides

What we learn from AI's biases



Mechanics (source: Pixabay)

In "How to Make a Racist AI Without Really Trying," Robyn Speer shows how to build a simple sentiment analysis system, using standard, well-known sources for word embeddings (GloVe and word2vec), and a widely used sentiment lexicon. Her program assigns "negative" sentiment to names and phrases associated with minorities, and "positive" sentiment to names and phrases associated with Europeans. Even a sentence like "Let's go get Mexican food" gets a much lower sentiment score than "Let's go get Italian food." That result isn't surprising, nor are Speer's conclusions: if you take a simplistic approach to sentiment analysis, you shouldn't be surprised when you get a program that embodies racist, discriminatory values. It's possible to minimize algorithmic racism (though possibly not eliminate it entirely), and Speer discusses several strategies for doing so.

I want to look at this problem the other way around. There's something important we can learn from this experiment, and from other examples of AI "gone wrong." AI never "goes wrong" on its own; all of our AI systems are built by humans, and reflect our values and histories.

O'Reilly AI Newsletter

Get the O'Reilly AI Newsletter

What does it mean when you build an AI system in the simplest possible way that you end up with a racially biased result? I don't think many AI developers would build such systems intentionally. I am willing to believe that many are naive and take free data sources at face value. That is exactly what is happening here: GloVe, a widely used collection of word embeddings, brings a lot of baggage with it, as does word2vec. But, just as programmers are more likely to be naive than evil, I don't think GloVe was built by people trying to perpetuate stereotypes. They just collected English language samples. They're a reflection of language as it is used.

All of which means we're facing a deeper problem. Yes, Speer's naive sentiment analysis is racist, but not because of the algorithm. It's because of the data; and not because the data is wrong, but because the data is right. The data wasn't collected with malice aforethought; it just reflects how we use language. Our use of language is full of racial biases, prejudices, and stereotypes. And while I would not recommend that anyone build and deploy a naive system, I appreciate examples like this because they hold up a mirror to our own usage. If we're willing to listen, they teach us about the biases in our own speech. They're metrics for our own poor performance.

Fairness is, by nature, aspirational: it's forward-looking. We want to be fair; we rarely look at the past and take pride in how fair we were. Data is always retrospective; you can't collect data from the future. Every datum we have reflects some aspect of the past, which means it almost always reflects history of prejudice and racism, both overt and covert. Our language is likely to be a better metric for our attitudes than any public opinion poll. Nobody *thinks* they are a racist; but our language says otherwise, and our algorithms reflect that.

We can (and we need to) analyze almost every example of algorithmic unfairness in this way. COMPAS, the tool for recommending bail and jail sentences, reflects a history of law enforcement that has fallen much more heavily on minorities. Minorities don't often get second chances; they don't get policemen who look the other way after saying "aw, he's basically a good kid" or "don't let me catch you doing that again." Poor urban neighborhoods get labeled "high risk zones," though if you look at a map of white collar crime, you'll see something much different. While COMPAS is a bad tool in the courtroom, it's an excellent tool for understanding the reality of how law enforcement works, and it's unfortunate it hasn't been used that way. (It might also be less unfair than predominantly white judges and juries, but that's another question.) Many of the problems around face recognition for dark-skinned people arise because cameras have long been designed to optimize for light skin tones. That's less a reflection on our technical capabilities than our cultural priorities. Amazon's initial same-day delivery service, which excluded heavily black and hispanic neighborhoods, doesn't reflect some evil intent; it reflects a long history of red-lining and other practices that forced minorities into ghettos. Exclusion jumped out of the data, and it's important to understand the histories that gave us that data.

When you get to the bottom of it, these aren't problems with the algorithms, or even with the data; they're problems with the ultimate source of the data, and that's our own actions. If we want better AI, we must be better people. And some of our bad AI could be the best tool we have for understanding how to be better people.

Article image: Mechanics (source: Pixabay).
